LA-UR-15-20907
Approved for public release; distribution is unlimited.

Title: Trinity Era Storage

Author(s): Lamb, Kyle E.

Intended for: Document release to Sandia National Laboratory.

Issued: 2015-02-09

Trinity Era Storage

Kyle Lamb

HPC-3 Infrastructure Team Lead

01/26/2015

Agenda

- Current Infrastructure
- Trinity requirements
- Limitations of existing solution
- Erasure code
- Trinity Campaign storage

Current Infrastructure

286TB Memory

lustre®

8PB
160GB/s Write

HPSS
High Performance Storage System

45PB
3.0GB/s Write
150MB/s Read

Current HPSS Archive Performance

HPSS Disk Performance

- 3.5 GB/s Write
  - Bottleneck HPSS disk cache or Metadata engine disk
- 1.3 GB/s Read
  - Bottleneck HPSS disk cache or Metadata engine disk

***Note**: Disk performance is using PSI with parallel movers, HSI work is needed to operate in parallel.*

HPSS Tape Performance

- 194MB/s *should be better
- Largest to date: 23TB file read back from tape with 4 tape drives
- Tape is bottleneck

***Caveat:** HPSS-only archive solution will require significant development to address n-to-n performance, RAIT, and design of a chunking utility.*

Current HPSS Archive Capacity

- HPSS growth rate: approximately 600TB per month sustained
  - 2X the memory footprint on the floor (Cielo, Luna, Typhoon, and Viewmaster2)
  - Growth rate has slowed, used to be 3X memory per month

HPSS contains 40PB of data in aggregate

- *The HPSS Disk cache has a capacity of 360TB*
- *Cache is written over twice every month.*

Most Labs are targeting 3-6 Months of disk cache vs. our 2 weeks

Proposed Trinity Archive Bandwidth to Disc Cache Requirements

- **ASSUMING CURRENT GOAL** of reading a checkpoint of 80% of main memory within 12 hours
- 2.0PB Memory * .8 = 1.6PB * 1024 TB/PB = 1638.4 TB
- 1638.4 TB * 1024 TB/GB = 1,677,721.6 GB
- 12 hrs * 60min/hr * 60sec/min = 43,200 seconds
- 1,677,72.6GB / 43,200 sec = **38.83 GB/sec Minimum**

Current HPSS Read Performance 198MB/s
* Given Above, Recall of 1.6PB would take 100 days, Assuming things don't break!

Data Set Size	Recall Time Window	Performance Required	Required Increase In performance
1.6PB	12hrs	38.8GB/s	200X
750TB	12hrs	17.8GB/s	92X
750TB	24hrs	8.9GB/s	46X
750TB	48hrs	4.5GB/s	23X

Archive at the scale of Cielo's current Parallel File System 40GB/s

Trinity Archive Capacity

Design Assumption: Expected growth rate of 2X main memory in operation

- 2PB for Trinity
- .5PB for CTS1, Luna, etc.
- 2.5PB * 2 = 5PB/month growth

Result: Expected usage of 5PB/month unless systematic usage policy changes are adopted e.g.,

Space quota (LLNL)
Recharge quota (SNL)

Note: A single 750TB file spans up to 88 T10KD tapes

By comparison: Current archive growth is 600TB/ month with Cielo and all other systems currently in operation

Largest file to date ever recalled from tape:
23TB @194MB/s
-> 30hours

A "baby" File in a Trinity world.

Caveat Detail: Required HPSS Development

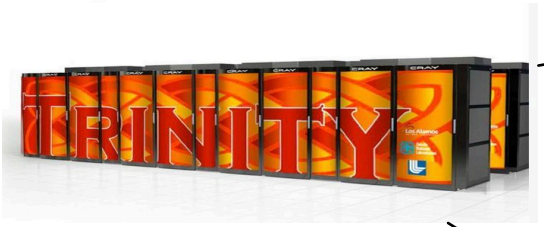
Note: Features are listed in priority order

- Transfer Agent
  - Multi-node transfers with HSI
- HPSS Metadata engine
  - Distributed database required (3.2M n-to-n files)
- RAIT(Redundant Array of Independent Tape)
  - Development underway to accommodate 14+2 RAIT
- HPSS Disk Cache
  - Increase disk cache to accommodate minimum of 2 months of data (10PB)
- HTAR
  - Enable Parallel HTAR utility (Ingest 3.2M n-to-n files)
- Chunk Utility
  - Required to enable multiple RAIT sessions

Tech Notes:

- Scale of HPSS Disk Cache changes if we augment archive
- Chunk Utility is only needed to scale beyond 2.1GB/s

Campaign Augmented Archive



2PB Memory

lustre®

82PB
1.45TB/s Write
2 File Systems

HPSS
High Performance Storage System

65PB
3.0GB/s Write
500MB/s Read

Burst Buffer

3.7PB
3.3TB/s Write

Campaign Storage

14-100s PB
50GB/s Write
40GB/s Read

Campaign Storage

A Technical solution that utilizes disk-based storage

- Leverages experience in recently completed “Campaign Storage” effort in open network
- Performance of system scales linearly with the amount of storage in use
- Utilizes Disk Scalable Units (DSUs) to provide large pools of storage on commodity (less expensive) disk
- Utilizes a relatively new storage algorithm referred to as erasure coding.

Caveats: Even with a Campaign-enhanced HPSS archive, it may be necessary to scale back the amount of data written to the archive (e.g. Quotas both in Campaign and HPSS).

Proposed Approach

- Build out a campaign enhanced storage solution to provide parallel file system performance with near archive level storage reliability
- Campaign storage becomes the long-term data storage location for large data sets and large n-to-n results
  - Not an unlimited resource: Quotas will be in place
- HPSS remains the archive location for high-value software repositories, important visualization files, important data sets, etc…
- Scaling requirements will necessitate a cap on files that can be transferred to HPSS, in the range of 30TB

What Disk-based Erasure-coded Storage Enables

- **Reliability:**
  - Creation of a higher parity set than is possible with RAID 6
  - Survival of 3-8 disk/partition failures without data loss
  - Highly fault tolerant parallel scalable storage
  - Much faster rebuilds when disk failures occur
- **Scalability:**
  - Scaling of large numbers of parallel disks that enables high bandwidth data transfers
  - Performance scales linearly with disk deployments
- **Flexibility:**
  - DSUs can be utilized with various file systems and storage solutions
  - GPFS, HPSS, TSM, etc…

Caveats Revisited:

- Quotas will probably be required even in a campaign-enhanced storage solution
  - Quotas are more flexible in a campaign storage solution
- The building blocks of Campaign storage scale linearly with both performance and capacity
  - Tape solutions offer the ability to scale capacity independent of the performance
- Development on HPSS is still required with a campaign-enhanced solution

Campaign-enhanced: Campaign storage does not replace HPSS

Campaign-enhanced solution allows us to scale performance more economically than HPSS alone

Performance and Capacity Campaign

Read Performance:

~4GB/S per Disk Scalable Unit (DSU)

$450K per DSU

Capacity:

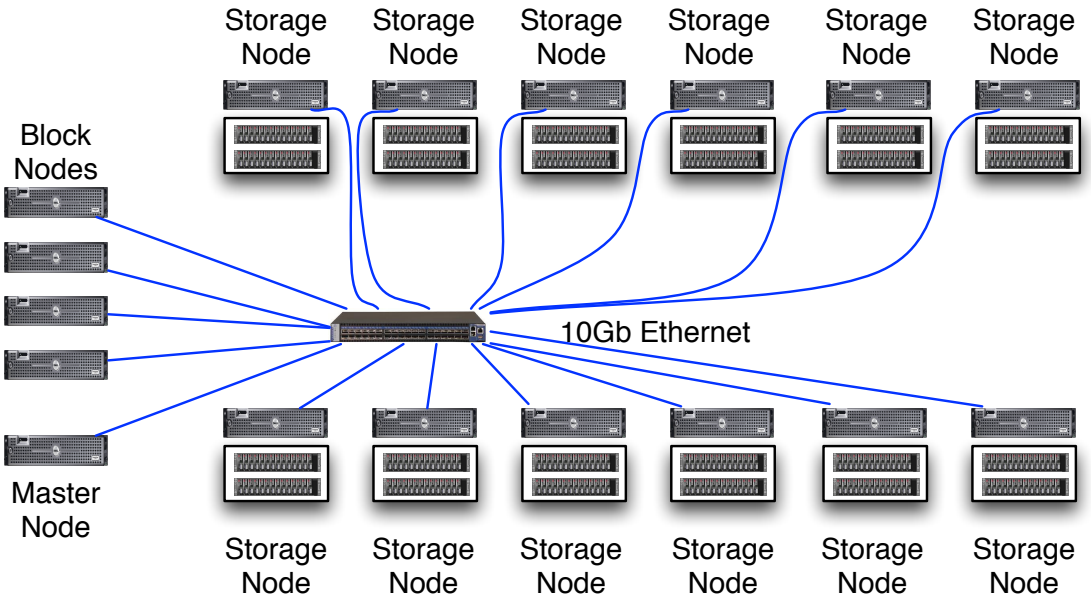
6.8PB per DSU usable (8TB drives)

*Additional capacity would be added by adding additional DSUs

Campaign Disk Scalable Unit

Disk Scalable Unit (DSU)

Xyratex Enclosure
84 Drives per Storage
Node 8.0PB RAW



Each DSU is configured with
a 40+8 parity set

UNCLASSIFIED

DSU Building Blocks

- Build out DSUs as large disk pools using a 40+8 parity stripe
- Can survive the loss of up to 8 disks/failure domains
  - If built with 48 servers we can survive the loss of 8 servers
- Up to 500MB/s per storage node
- Goal to attain 500MB/s to 1GB/s per GPFS connector node
- Testing with DSU behind HPSS and TSM

Note:
- Erasure Code parity is variable
- Reliability calculator shows 14 9's data reliability with 40+8

HPSS use cases:
Large file landing Zone
2nd copy pool small files

TSM use case:
2nd copy on disk
Customer archive solutions

Campaign Storage for Trinity

- Deploy ~30PB this year
- Target 1GB/s per PB of storage
- Utilize FTAs that will interface with Lustre FS from Trinity
- FTAs will allow access to both Campaign and HPSS
- FTAs will be utilized for local and remote data transfer
- PFTOOL utilized for local data transfer HSI, pftp, etc. for remote

Tech Notes:

- Performance scales with storage (100GB/s eventual target
- FTAs and Lustre single client performance may be bottleneck
- FTAs will utilize IPOIB for access to HPSS, Campaign, and Lustre

Questions?